Master of Science in Business Analytics Fall 2024

BA 882: Deploying Analytics Pipelines

Instructor: Brock Tibert
Office: 644A
Class: Tuesday 3:30-6:15
Office Hours: Hour before class
or by appointment
Room: 419
e-mail: btibert@bu.edu

COURSE DESCRIPTION & LEARNING GOALS

This course was designed to address the foundational skills necessary to move data analysis tasks to the cloud to support production workloads. A common pattern is for analysts to write code, build models on their laptops, and then hand off that work for other team members to deploy. This offering will extend the skills taught in the program, illustrating the process of creating and deploying data and machine learning pipelines in the cloud. Moreover, this course will discuss the responsibilities data teams are now facing with the rise of generative AI, including the emerging patterns to deploy the data pipelines required to support LLM-based tasks for an organization.

As noted above, careful consideration was given to the required courses of this program when constructing this offering. The skills to work with and manage data learned in BA775 and BA780 will be leveraged to construct data integration pipelines. The machine learning methods discussed in BA810 and BA820 will be revisited when discussing how data teams support and deploy machine learning models as APIs for use in downstream products and applications.

Upon completion of this course, you will have learned:

- **Data Ops**: Compare and contrast various data integration strategies such as ETL, ELT, and reverse ETL, as well as how to implement monitoring, observability, and data quality controls while building your cloud-based data pipelines for downstream tasks. This will include how to take your local data workflows to the cloud to demonstrate how you can move from development to "production" via the command line, git, docker, Infrastructure as Code, and various cloud services and providers.
- **Cloud Data Warehousing, Storage, and BI:** Analytics-focused warehouses and data lakes, including NoSQL and vector databases. In addition, you will learn how to take the modeled data and expose this to data consumers via standard BI tools as well as how to automate reports in a variety of formats (PDF, slide decks, etc.).
- **ML-Ops**: Build pipelines to train, evaluate, and deploy ML models via cloud services, including inference for use in a variety of applications.

- **LLM-Ops:** You will learn how to contrast the ETL/ELT strategies learned earlier in the semester with the emerging patterns required to build end-to-end analytics solutions that are backed by generative AI.

PREREQUISITES

Below are the following courses that are required to be completed prior to enrolling in this offering.

- QST BA 775, QST BA 780, QST BA 810, QST BA 820

If the requirements above are not satisfied, you may also be able to register for this course with the approval of the instructor.

COURSE STRUCTURE & PEDAGOGY

This course was designed to promote active learning via hands-on lectures and coursework. Each session will include about an hour of lecture/discussion/QA/debate, and the last two hours of active learning.

- Pre-recorded lectures before class discussions and lectures
- In-class discussion via hands-on work with in-class exercises
- Team-based Micro-Projects to frame each learning objective and the framing for the next test
- Individual problem sets and quizzes that evaluate and assess both conceptual and technical understanding

COURSE MATERIALS

The following tools will be used during this course:

- Github Account and Github Desktop - Account created here, Desktop installation here
- Docker Desktop
- Visual Studio Code - Install the appropriate version for your operating system here. You are free to use any IDE you prefer, but I will be using VS Code and its ecosystem in class.
- Miniconda - Python management, Install is located here. Anaconda Distribution should also enable the necessary command line tool.
- AWS|GCP Cloud Accounts - More details to be supplied on Blackboard
- DataCamp for Supplemental Practice and Review - Signup link here
- Arkaiv - Please review the Attendance App document on Blackboard to get started

While there isn't a primary textbook for this course, the following are excellent resources when learning about deploying data pipelines, machine learning models, and leveraging generative AI tools within a variety of analytics tasks. The links below are freely available online and can be used as supplemental references for this course, as well as other offerings within this program.

- Learning SQL (freely available via the BPL and O'Reillly)
- Fundamentals of Data Engineering (freely available via the BPL and O'Reillly)
- Data Wrangling on AWS (freely available via the BPL and O'Reillly)
- Analytics Engineering with SQL and dbt (freely available via the BPL and O'Reillly)

- Practical Machine Learning with AWS ([freely available via the BU Library](#))
- Google Cloud Platform for Data Science ([freely available via the BU Library](#))

The Boston Public Library's eCard provides you with access to Oreilly's Learning Platform. After receiving your membership card, the O'Reilly books listed above can be found via a search on the BPL website. This will redirect you to the O'Reilly learning platform. You can register for your BPL eCard using [this link](#).

ARTIFICIAL INTELLIGENCE (AI) POLICY

If used correctly, generative AI systems (like ChatGPT) can serve as powerful tools for learning and idea refinement. In this course, I encourage you to leverage generative AI systems to explore concepts iteratively through conversation, much like you would engage with a peer, TA, or instructor. You can use AI assistants to help generate ideas, brainstorm, and reason about the concepts covered in this course. However, you should note that the material generated by these programs may be incomplete, problematic, or even completely wrong. Be aware that relying on AI may also stifle your independent thinking and creativity.

I believe the intellectual growth you gain from working through a difficult problem and discovering the answer for yourself cannot be replicated by simply reading a pre-generated answer. After careful consideration, the course policies around the use of AI assistants are as follows:

- Unless explicitly stated, the use of AI assistants is prohibited on take-home assessments like quizzes and tests.
- The use of AI assistants for your team projects is encouraged. If you decide to use these tools, it is your responsibility to fact-check the insights you gain and perform your due diligence. You are accountable for all work submitted and will be asked to include an appendix documenting the prompts and outputs that supported your work. This appendix must also include appropriate information about the assistant used (which tool, version, etc.).

While AI can provide instant feedback, it is important to remember that it is a machine, not a human expert. It is incapable of independent thinking or making judgments based on personal experiences, cultural contexts, or ethical considerations. Therefore, if used, AI should be employed as a complementary tool to academic work, not as a replacement for your thinking and analysis.

The use of AI when prohibited will be considered academic dishonesty and handled accordingly.

DIVERSITY AND INCLUSION

- In developing this program, we have aimed to be thoughtful about how identity and culture impact the program content.
- We invite you to share your personal experiences and perspectives related to the program content. If there are topics of conversation that you feel would benefit from the incorporation of social context, a differing perspective, or Questrom's Office of Diversity & Inclusion, please inform us and we will explore resources and opportunities for us to engage a wide variety of perspectives in our classrooms.

PROGRAM POLICIES

Classroom Conduct
- Professionalism: Students are expected to follow Boston University's Student Codes of Conduct. They are expected to conduct themselves with respect and professionalism and can expect the same treatment. A part of the training is learning what is expected in a workplace, and we try to enforce those standards throughout the program.
- Punctuality: Students must arrive at the classes and meetings on time.
- Name Tents: Students are expected to keep the name tents in front of them during the class. This helps peers and faculty know them faster.
- Participation: Students are expected and strongly encouraged to ask questions and get involved in class discussions. This helps everyone get more out of the class.
- Cell Phone: Students cannot use their cell phones during class or exams unless specifically instructed. If you need to take an urgent call, simply leave the classroom and return as soon as possible.
- Activities Unrelated to Class: Activities unrelated to the class are not allowed during the class. These include, but are not limited to, the use of social media, news sites, online video sites, gaming, email checking, and/or writing unless explicitly asked by the instructor for the class.
- Absence Policy: If you must miss a class, notify your instructor before the class.
- Honor Code: All students must read and abide by the Academic Conduct Code. If you are uncertain about which types of collaborations are allowed in the class, check with the instructor.

Grading
Questrom School of Business follows program-wide guidelines for MS-BA courses that are graded:

- No more than 50% of grades in the class should be in the A or A- range.
- Grades in the C+, C, or C- range are legitimate passing grades, and will be used in the case of very low-performing students.

If you have any questions about grades that you receive on particular assignments, you must raise them within two weeks of receiving your grade on that assignment. Unless we have made computational errors, we will be unable to alter grades after final grades have been determined. If you have particular grade-related considerations that you think are important, please raise these with your instructor as early as possible (during the first half of the semester at the latest!), so that your instructor can help you approach the course in a way that will help you achieve your best possible performance.

If, after discussing your grade with the instructor, your grading concerns have not been addressed to your satisfaction, contact the MS-BA Faculty Director.

Accommodations for Students with Special Needs
In keeping with University policy, any student with a disability who needs or thinks they need academic accommodations must call the Office of Disability Services at 353-3658 or stop by 19 Deerfield Street to arrange a confidential appointment with a Disability Services staff member. Accommodation letters must be delivered to the Faculty Directory in a timely fashion (within two weeks of the date on the letter and not later than two weeks before any major examination). Please note that accommodations will not be delivered absent an official letter of accommodation.

Sexual Misconduct/Title IX Policy

The Questrom School of Business is committed to fostering a safe learning environment for all members of the community and preventing sexual misconduct. All forms of sexual misconduct, including rape, acquaintance rape, sexual assault, domestic and dating violence, stalking, and sexual harassment are violations of Boston University's policies, whether they happen on campus or off campus. Title IX of the Education Amendments of 1972 is a federal civil rights law that prohibits sex-based discrimination in federally funded education programs and activities. This law makes it clear that violence and harassment based on sex and gender is a Civil Rights offense subject to the same kinds of accountability and the same kinds of support applied to offenses against other protected categories such as race, national origin, etc. If you or someone you know has been harassed or assaulted, you can find the appropriate resources at http://www.bu.edu/safety/sexual-misconduct/.

COURSE RECORDING POLICY

All class sessions will be recorded for the benefit of registered students who are unable to attend live sessions (either in person or remotely) due to time zone differences, illness, or other special circumstances. Recorded sessions will be made available to registered students ONLY via their password-protected Questrom Tools account. Students may not share such sessions with anyone not registered in the course and may certainly not repost them in a public platform. Students have the right to opt out of being part of the class recording. Please contact your instructor or teaching assistant to discuss options for attending the course in such cases.

ATTENDANCE

Satisfactory class contributions require attendance at every class session; preparation of all materials for every session; and active, quality participation in class discussions. Simply attending class, however, does not constitute a positive contribution to the class and will not yield high-class contribution scores. Recognizing that you are facing complex demands on your time, we can excuse two absences during the term (for any reason). Students who miss 3 or 4 sessions will lose 3 and 6 points on their final course grade, respectively. Students who miss more than 5 sessions will have their final grades reduced an entire letter grade and students who miss more than 7 sessions may fail the course as a result. Assignments are always due based on the stated deadline within Blackboard, even if you are unable to attend class that day.

COURSE EVALUATION AND EXPECTATIONS

The following evidence will be used to determine your grade:

Grading Component	Percent of Course Grade
Quizzes	20%
Midterm	15%
Team Projects	40%
Final	15%
Class Contribution	10%
Total	**100%**

Quizzes: (20%)

Throughout the course, there will be a series of quizzes designed to reinforce key concepts and ensure understanding of the material covered during class discussions and readings. Each quiz will be short and focused, covering recent topics to encourage consistent engagement with the content. The quizzes will be conducted online and may include a mix of multiple-choice, short-answer, and practical problem-solving

questions. Together, these quizzes will make up 20% of your final grade, providing a steady check on your progress and understanding throughout the course.

Course Contributions (10%):
For this course I expect students to actively participate by engaging in discussions, sharing insights from their experiences, and critically analyzing real-world data. Students are encouraged to ask questions, challenge assumptions, and bring diverse perspectives to enhance collaborative learning. Participation in-class exercises, thoughtful input during group discussions, and constructive feedback to peers will be key components of your contributions.

Midterm and Final Exams (20% each):
The two exams will be based on class notes, in-class discussions and exercises, and the assigned readings and cases. Anything discussed in class can be tested in these unless explicitly excluded. Therefore, being present and participating in class is critical for success on the exams. The exam format will be mostly short-answer questions aimed at assessing analytical reasoning as it relates to the topics and methodologies covered.

Team Project: (30%)

In teams of 4-5, you will collectively define a business problem and identify real-world data sources to use throughout the semester. Your team will develop a data pipeline to populate a cloud data warehouse, deploy a machine learning model using this data, and enrich the data through generative AI tools and pipelines in the final phase of the course. In class, we will explore various publicly available data sources and repositories that your team may choose from, but it is important to note that the data feeds chosen by your team need to be updated no less than weekly.

Your project grade will be based on (a) the novelty and relevance of the application you are designing to solve your stated business problem, (b) the rigor and attention to detail in your work, and (c) the organization of your project. Each member of your team will have assigned roles and responsibilities, and you will keep track of your project on Github based on methods discussed in class. How you manage your project and collaborate using the platform will also be evaluated.

After each, phase of the course, your team will present to the class the progress you have made, and perform any demonstrations you deem appropriate, including how to use reports and interfaces you make available for your peers to interact with. Your team's demonstrations after each phase will also include a peer evaluation component. Students from another team will act as stakeholders, asking questions about your project, your progress, and if applicable, their experience interfacing with apps or services you provide. They will submit evaluation forms based on your presentation, work, and responses to their questions.

Your final project grade will be determined by the peer evaluation (30%), the instructor's evaluation of the team (40%), and your teammates' assessment (30%). Acting as a stakeholder for other teams is also part of your grade, and failure to participate will negatively impact your overall score.

COURSE SCHEDULE AND CLASS LAYOUT

<div align="center">

Class Schedule
I reserve the right to change the syllabus at any time.

</div>

Session	Day	Date	Learning Objectives	Additional Notes/Recommended Materials
Data Ops: ETL/ELT and Warehousing				
1	T	9/3	- Course intro - The Modern Data Stack - Local dev tools and Cloud discussion - Crash review: python, git, cli, docker - Brief discussion of managing technical project workflows via Github projects - Brief Recap of SQL for data analysis	● Datacamp (DC): Exploratory Data Analysis ● DC: git, docker, and GCP courses as needed ● Review local machine setup **Teams will be assigned after class 1**
2	T	9/10	- Deep Dive into patterns for data pipelines - ETL vs ELT considerations - Data Modeling/ERD - SQL (DDL/DML)	● Datacamp: Data Warehousing Concepts 1-3
3	T	9/17	- Cloud Data Warehousing Considerations - Pipeline examples for data modeling - Stand up the start cloud data warehouse (design) - Get a pipeline working locally/start development considerations	● Datacamp: Data Warehousing Concepts 4 **Quiz 1 outside of class**
4	T	9/24	- Pipeline Development/Orchestration - QA - Data checks within the pipeline - Deploy the pipeline via the web console or CLI - Monitoring / Observability / Alerting considerations	● Datacamp: Intro to Airflow with Python ● Datacamp Tutorial: Prefect
5	T	10/1	- Reporting Automation - BI tools (Superset, Redash) - Data Products (e.g. Streamlit)	● Getting started with Quarto ● Superset ● Streamlit Gallery
6	T	10/8	- Packaging your projects - From point-and-click to IaC deployments - Step back and extra time on core themes if necessary	**Quiz 2 Outside of Class**
7	T	10/22	- Team Project 1 Presentation - Section Recap and where to go next	
Midterm Takehome - Details to be discussed in Class and on Blackboard				
ML Ops				

8	T	10/29	- Brief 810 recap - ML vs DL Considerations when training and deploying in the cloud for production - Cloud-based tools to facilitate Model development	● AWS: Chapters 4 and 5 ● GCP: Chapter 2
9	T	11/5	- ML Deployment: From model fit to Production inference	● AWS: Chapter 6 ● GCP: Chapter 3 **Quiz 3 outside of class**
10	T	11/12	- Team Project Presentations	
			LLM Ops	
11	T	11/19	- Brief 820 recap - Data ingestion/Knowledge Graph/Vector via NoSQL data stores - Infrastructure setup and data modeling considerations - Data pipeline completed to power generative-ai-backed analytics solutions	https://arxiv.org/pdf/2402.06196.pdf Podcast: High Performance Generative AI Applications with Ram Sriharsha, CTO at Pinecone
12	T	11/26	- Fine-tuning versus Retrieval Augmented approaches (RAG) - Deploy a simple RAG pipeline	**Quiz 4 outside of class**
13	T	12/3	- RAG continued - Pipeline evaluation metrics and how to finetune the pipeline - What are Agents, and how do they compare to earlier promises made by "chatbot" solutions - Emerging patterns used in generative AI analytics applications (e.g. re-ranking). - A discussion of emerging frameworks, methods, and best practices	
14	T	12/10	Team Project Presentations	
		12/18	Registrar Set Date: Final 3-5pm	